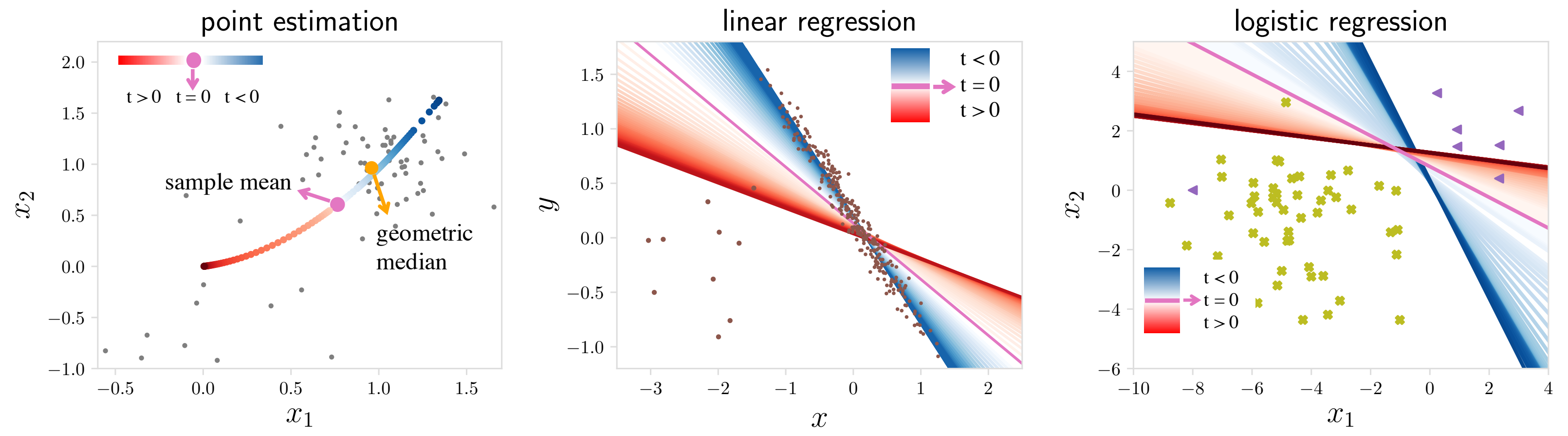
Tilted Empirical Risk Minimization

Tian Li* (CMU), Ahmad Beirami* (Facebook AI), Maziar Sanjabi (Facebook AI), Virginia Smith (CMU)

Tilted ERM (TERM) Objective

Address the deficiencies of ERM $\min_{\theta} R(\theta) := \frac{1}{n}\sum_{i=1}^{n} f(x_i;\theta)$ in a unified framework:

Tilted ERM: $\min_{\theta} \tilde{R}(t;\theta) := \frac{1}{t}\log\left(\frac{1}{n}\sum_{i=1}^{n} e^{tf(x_i;\theta)}\right)$ ← use the tilt parameter t to tune the impact of individual losses



$f(x;\theta)=\frac{1}{2}\theta^\top(\theta-2x)$ $f(x,y;\theta)=\frac{1}{2}(\theta^\top x-y)^2$ $f(x,y;\theta)=\log(1+e^{-y\theta^\top x})$

TERM:

- increases or decreases the influence of outliers to enable fairness or robustness
- can be viewed as a smooth approximation to quantile losses
- can be solved efficiently with batch and stochastic optimization methods
- can be used for a multitude of applications, achieving competitive with existing solutions tailored to these individual problems, and enable entirely new applications

TERM Solver

Batch case (hierarchical tilting)

$\tilde{R}_g$ ← tilted loss on group g, $w_g \leftarrow \frac{|g|e^{t\tilde{R}_g}}{\sum_{g\in[G]}|g|e^{t\tilde{R}_g}}$, $\theta \leftarrow \theta - \alpha\sum_{g\in[G]} w_g \nabla_\theta \tilde{R}_g$

✓ convergence rate scales linearly with t

Stochastic case (hierarchical tilting)

sample a group g from a Gumbel-Softmax distribution with based on $\tilde{R}_g$,

$\tilde{R}_{g,B}$ ← tilted loss on a mibi-batch B in group g,

$e^{t\tilde{R}_g} \leftarrow (1-\lambda)e^{t\tilde{R}_g} + \lambda e^{t\tilde{R}_{g,B}}$, ← estimate the weight normalizer

use $e^{t\tilde{R}_g}$ to update weights and θ

★ *TERM can be used in sample-level, group-level, and hierarchical tilting, running time within 2x of ERM*

Properties

Re-weighting samples to magnify/suppress outliers

$$\tilde{R}(t;\theta) = \frac{1}{t}\log\left(\frac{1}{n}\sum_{i=1}^{n} e^{tf(x_i;\theta)}\right)$$

gradients: $\nabla_\theta \tilde{R} = \sum_{i=1}^{N} w_i(t;\theta)\nabla_\theta f(x_i;\theta)$, and $w_i(t;\theta) = \frac{e^{tf(x_i;\theta)}}{\sum_{j\in[N]} e^{tf(x_j;\theta)}}$

Trade-off between average loss and max-/min-loss

as t moves from 0 to $+\infty$, the **average loss** will increase, and the **max-loss** will decrease

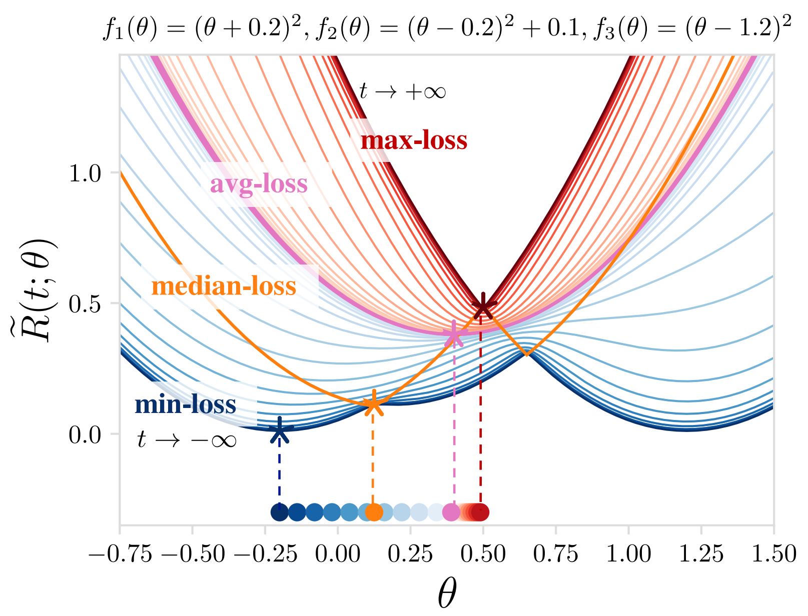
as t moves from 0 to $-\infty$, the **average loss** will increase, and the **min-loss** will decrease

[Empirical bias-variance tradeoff] as t increases, the **average loss** will increase, and the **loss variance** will decrease => better generalization

Approximation of quantile losses

quantile losses: $\arg\min_\theta \; Q(a;\theta) := \frac{1}{N}\sum_{i\in[N]} \mathbb{I}\{f(x_i;\theta)\geq a\}$

quantile loss solutions can be approximated by TERM solutions



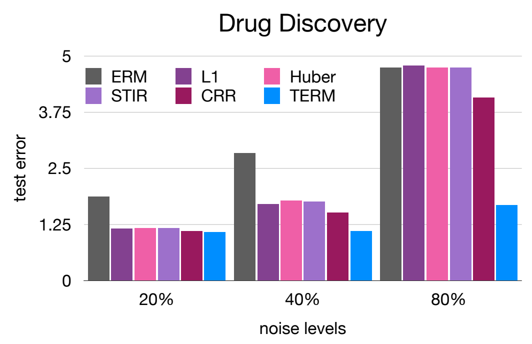
TERM objectives for a squared loss problem with N=3. Tilted losses recover min-loss, avg-loss, and max-loss. TERM is smooth for all finite t and convex for positive t. TERM solutions approximate median-loss minimizer.

See paper for complete theoretical results

Applications

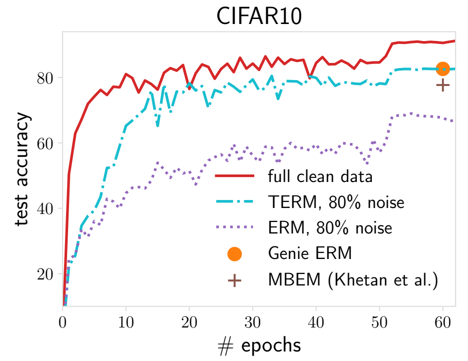
On real-world ML applications, TERM is superior than (or competitive with) existing, problem-specific state-of-the-art solutions

Robust regression



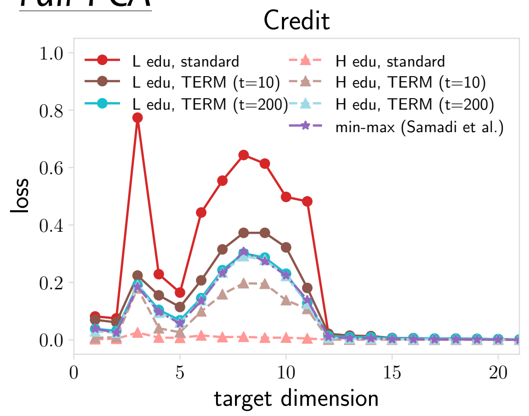
TERM is competitive with robust regression baselines, particularly in high noise regimes.

Robust classification



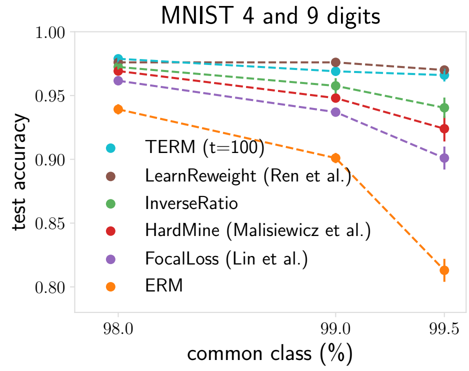
TERM completely removes the impact of noisy annotators.

Fair PCA



TERM-PCA flexibly trades the performance on the high (H) edu group for the performance on the low (L) edu group.

Handling class imbalance



TERM is competitive with state-of-the-art methods for classification with imbalanced classes.

Objectives	imbalance, clean		imbalance, noisy	
	minority	overall	minority	overall
ERM	0.503	0.888	0.240	0.831
GCE	0.503	0.888	0.324	0.849
LearnReweight	0.800	0.904	0.532	0.856
RobustRegRisk	0.622	0.906	0.051	0.792
FocalLoss	0.806	0.918	0.565	0.890
TERM	**0.836**	**0.924**	**0.806**	**0.901**

TERM is able to handle compound issues, e.g., the existence of noisy samples and imbalanced classes

see paper for all results

Future Work

- Other applications of the TERM framework (e.g., meta-learning, GAN training)
- Other properties of TERM (e.g., adversarial robustness)
- Generalization of the TERM objective with respect to t
- Further connections with other risks (DRO, Conditional Value-at-Risk, Invariant Risk Minimization, etc)

Code: https://github.com/litian96/TERM